2.2 Punktschätzung

Ziel: Finde ein möglichst gutes Schätzverfahren und damit einen möglichst guten Schätzwert für eine bestimmte Kenngröße ϑ (Parameter) der Grundgesamtheit, z.B. den wahren Anteil der rot/grün-Wähler, den wahren Mittelwert, die wahre Varianz, aber auch z.B. das wahre Maximum (z.B. von Windgeschwindigkeit).

2.2.1 Schätzfunktionen

Gegeben sei die in Kapitel 2.1 beschriebene Situation, also eine i.i.d. Stichprobe $X_1, \ldots, X_n$ eines Merkmales $\tilde{X}$.

Definition 2.1.

Sei $X_1, \ldots, X_n$ i.i.d. Stichprobe. “Jede“ Funktion

$$T = g(X_1, \ldots, X_n)$$

heißt $Schätzer$ oder $Schätzfunktion$.

Inhaltlich ist $g(\cdot)$ eine Auswertungsregel der Stichprobe: „Welche Werte sich auch in der Stichprobe ergeben, ich wende das durch $g(\cdot)$ beschriebene Verfahren auf sie an. (z.B. ich bilde den Mittelwert der Daten)“

Typische Beispiele für Schätzfunktionen:

1. Arithmetisches Mittel der Stichprobe:

$$\bar{X} = g(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

   Für binäre X_i mit $X_i \in \{0, 1\}$ ist $\bar{X}$ auch die relative Häufigkeit des Auftretens von „$X_i = 1$“ in der Stichprobe

2. Stichprobenvarianz:

$$\tilde{S}^2 = g(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - (\bar{X})^2 = \frac{1}{n}\left(\sum X_i^2 - n\bar{X}^2\right)$$

3. Korrigierte Stichprobenvarianz:

$$S^2 = g(X_1, \ldots, X_n) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n \cdot \bar{X}^2 \right)$$

4. Größter Stichprobenwert:

$$X_{(n)} = g(X_1, \ldots, X_n) = \max_{i=1,\ldots,n} X_i$$

5. Kleinster Stichprobenwert:

$$X_{(1)} = g(X_1, \ldots, X_n) = \min_{i=1,\ldots,n} X_i$$

Schätzfunktion und Schätzwert: Da $X_1, \ldots, X_n$ zufällig sind, ist auch die Schätzfunktion $T = g(X_1, \ldots, X_n)$ *zufällig*. Zieht man mehrere Stichproben, so erhält man jeweils andere Realisationen von $X_1, \ldots, X_n$, und damit auch von T.

Die Realisation t (konkreter Wert) der Zufallsvariable T (Variable) heißt *Schätzwert*.

Man hat in der Praxis meist nur eine konkrete Stichprobe und damit auch nur einen konkreten Wert t von T. Zur Beurteilung der mathematischen Eigenschaften werden aber alle denkbaren Stichproben und die zugehörigen Realisationen der Schätzfunktion T herangezogen.

D.h. beurteilt wird *nicht* der einzelne Schätzwert als solcher, sondern die Schätzfunktion, als *Methode*, d.h. als *Regel* zur Berechnung des Schätzwerts aus der Stichprobe.

Andere Notation in der Literatur: $\hat{\vartheta}$ Schätzer für ϑ.

Dabei wird nicht mehr direkt unterschieden zwischen Zufallsvariable (bei uns Großbuchstaben) und Realisation (bei uns klein). $\Longrightarrow$ Schreibe $\hat{\vartheta}(X_1, \ldots, X_n)$ bzw. $\hat{\vartheta}(x_1, \ldots, x_n)$ wenn die Unterscheidung benötigt wird.

Bsp. 2.2.

Durchschnittliche Anzahl der Statistikbücher in einer Grundgesamtheit von Studierenden schätzen.

- Grundgesamtheit: Drei Personen $\tilde{\Omega} = \{\tilde{\omega_1}, \tilde{\omega_2}, \tilde{\omega_3}\}$.
- Merkmal $\tilde{X}$: Anzahl der Statistikbücher

$$\tilde{X}(\tilde{\omega}_1) = 3 \quad \tilde{X}(\tilde{\omega}_2) = 1 \quad \tilde{X}(\tilde{\omega}_3) = 2.$$

  Wahrer Durchschnittswert: $\mu = 2$.

- Stichprobe X_1, X_2 ohne Zurücklegen (Stichprobenumfang $n = 2$):

$$X_1 = \tilde{X}(\omega_1) \qquad X_2 = \tilde{X}(\omega_2)$$

  wobei

$$\omega_1 \text{ erste gezogene Person, } \omega_2 \text{ zweite gezogene Person.}$$

Betrachte folgende möglichen Schätzer:

$$
\begin{aligned}
T_1 &= g_1(X_1, X_2) = \bar{X} = \frac{X_1 + X_2}{2} \\
T_2 &= X_1 \\
T_3 &= g(X_1, X_2) = \frac{2}{3} X_{(2)} = \frac{2}{3} \max(X_1, X_2)
\end{aligned}
$$

2.2.2 Gütekriterien

Beurteile die Schätzfunktionen, also das Verfahren *an sich*, nicht den einzelnen Schätzwert. Besonders bei komplexeren Schätzproblemen sind klar festgelegte Güteeigenschaften wichtig.

Natürlich ist auch zu Beginn genau festzulegen, was geschätzt werden soll. Im Folgenden sei der Parameter ϑ stets eine eindimensionale Kenngröße der Grundgesamtheit (z.B. Mittelwert, Varianz, Maximum)

Der Punkt ist, dass T zufällig ist; der Wert schwankt mit der konkreten Stichprobe.

- Man kann also nicht erwarten, dass man immer den richtigen Wert trifft.
- Die Beurteilung der Güte des Schätzers bezieht sich auf Kenngrößen seiner Verteilung (v.a. Erwartungswert und Varianz)

Definition 2.3. *(Erwartungstreue, Bias)*

Gegeben sei eine Stichprobe $X_1, \ldots, X_n$ und eine Schätzfunktion $T = g(X_1, \ldots, X_n)$ (mit existierendem Erwartungswert).

- T heißt *erwartungstreu für den Parameter* ϑ, falls gilt

$$\mathrm{E}_\vartheta(T) = \vartheta$$

  für alle ϑ.

- Die Größe

$$\mathrm{Bias}_\vartheta(T) = \mathrm{E}_\vartheta(T) - \vartheta$$

  heißt *Bias* (oder *Verzerrung*) der Schätzfunktion.

Man schreibt $\mathrm{E}_\vartheta(T)$ und $\mathrm{Bias}_\vartheta(T)$, um deutlich zu machen, dass die Größen von dem wahren ϑ abhängen. Erwartungstreue Schätzfunktionen haben per Definition einen Bias von 0.

Anschauliche Interpretation:

Bsp. 2.4. [Fortsetzung des Beispiels]

Nehmen Sie an, die Stichprobenziehung sei gemäß einer reinen Zufallsauswahl erfolgt, d.h. jede Stichprobe hat dieselbe Wahrscheinlichkeit gezogen zu werden (hier dann $\frac{1}{6}$). Sind die oben betrachteten Schätzfunktionen T_1, T_2, T_3 erwartungstreu?

Zur Beurteilung: Alle möglichen Stichproben (vom Umfang $n = 2$, ohne Zurücklegen) betrachten:

Nummer der Stichprobe	Personen in der Stichprobe	Realisationen von				
		X_1	X_2	T_1	T_2	T_3
1	$\tilde{\omega}_1$, $\tilde{\omega}_2$	3	1	2	3	2
2	$\tilde{\omega}_1$, $\tilde{\omega}_3$	3	2	2.5	3	2
3	$\tilde{\omega}_2$, $\tilde{\omega}_1$	1	3	2	1	2
4	$\tilde{\omega}_2$, $\tilde{\omega}_3$	1	2	1.5	1	$1.\bar{3}$
5	$\tilde{\omega}_3$, $\tilde{\omega}_1$	2	3	2.5	2	2
6	$\tilde{\omega}_3$, $\tilde{\omega}_2$	2	1	1.5	2	$1.\bar{3}$

Verteilung von T_1, T_2, T_3 siehe bei Erwartungstreue

Für die Träger $\mathcal{T}_i$ von $T_i,\ i = 1, 2, 3$ gilt:

$$\begin{aligned}
\mathcal{T}_1 &= \{1.5, 2, 2.5\} \\
\mathcal{T}_2 &= \{1, 2, 3\} \\
\mathcal{T}_3 &= \{1.\bar{3}, 2\}
\end{aligned}$$

Bei T_1 gilt: $\quad P(\{T_1 = 1.5\}) = P(\{T_1 = 2\}) = P(\{T_1 = 2.5\}) = \frac{2}{6} = \frac{1}{3}$

Bei T_2 gilt: $\quad P(\{T_2 = 1\}) = P(\{T_2 = 2\}) = P(\{T_2 = 3\}) = \frac{2}{6} = \frac{1}{3}$

Bei T_3 gilt: $\quad P(\{T_3 = 1.5\}) = \frac{2}{6} = \frac{1}{3}; \qquad P(\{T_3 = 3\}) = \frac{4}{6} = \frac{2}{3}$

und damit bei $\vartheta = \mu = 2$:

- $\mathbb{E}_2(T_1) = \sum_{t_1 \in \mathcal{T}_1} t_1 \cdot P(\{T_1 = t_1\}) = \frac{1}{3}(1.5 + 2 + 2.5) = 2.$

  In der Tat gilt allgemein (vgl. die nächste Bemerkung): Das arithmetische Mittel ist erwartungstreu für den Erwartungswert.

$$\mathbb{E}_2(T_2) = \sum_{t_2 \in \mathcal{T}_2} t_2 \cdot P(\{T_2 = t_2\}) = \frac{1}{3}(1 + 2 + 3) = 2$$

  Wieder gilt allgemein: Einzelne Stichprobenvariablen ergeben auch erwartungstreue Schätzer für den Erwartungswert, sind aber weniger geeignet (s.u.).

$$\mathbb{E}_2(T_3) = \sum_{t_3 \in \mathcal{T}_3} t_3 \cdot P(\{T_3 = t_3\}) = \frac{1}{3} \cdot 1.\bar{3} + \frac{2}{3} \cdot 2 = \frac{16}{9} \neq 2$$

  T_3 ist also nicht erwartungstreu. Es gilt

$$\mathrm{Bias}(T_3) = \mathbb{E}_2(T_3) - 2 = \frac{16}{9} - \frac{18}{9} = -\frac{2}{9}$$

Bem. 2.5. [Bias und Erwartungstreue bei einigen typischen Schätzfunktionen]

- Das arithmetische Mittel $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist erwartungstreu für den Mittelwert μ einer Grundgesamtheit: Aus $X_1, \ldots, X_n$ i.i.d. und $\mathrm{E}_\mu(X_1) = \mathrm{E}_\mu(X_2) = \ldots = \mu$ folgt:

$$
\begin{aligned}
\mathrm{E}(\bar{X}) &= \mathrm{E}_\mu\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\,\mathrm{E}_\mu\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\cdot n \cdot \mu = \mu
\end{aligned}
$$

- Sei σ^2 die Varianz in der Grundgesamtheit. Es gilt

$$\mathrm{E}_{\sigma^2}(\tilde{S}^2) = \frac{n-1}{n}\sigma^2,$$

also ist $\tilde{S}^2$ *nicht* erwartungstreu für σ^2.

$$\mathrm{Bias}_{\sigma^2}(\tilde{S}^2) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

(Für $n \to \infty$ geht $\mathrm{Bias}_{\sigma^2}(\tilde{S}^2)$ gegen 0, $\tilde{S}^2$ ist „*asymptotisch erwartungstreu*“.)

- Für die korrigierte Stichprobenvarianz gilt dagegen:

$$\begin{aligned} \mathrm{E}_{\sigma^2}(S^2) &= \mathrm{E}_{\sigma^2}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) \\ &= \mathrm{E}_{\sigma^2}\left(\frac{1}{n-1}\cdot\frac{n}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) \\ &= \mathrm{E}_{\sigma^2}\left(\frac{n}{n-1}\tilde{S}^2\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2 \end{aligned}$$

  Also ist S^2 erwartungstreu für σ^2. Diese Eigenschaft ist auch die Motivation dafür, von einer Korrektur der Stichprobenvarianz zu sprechen.

- *Vorsicht:* Im Allgemeinen gilt, wie in Kapitel 1.5.3 ausgeführt, für beliebige, nichtlineare Funktionen g

$$\mathrm{E}\,g(X) \neq g(\mathrm{E}(X)).$$

  Man kann also nicht einfach z.B. $\sqrt{\cdot}$ und E vertauschen. In der Tat gilt: S^2 ist zwar erwartungstreu für σ^2, aber $\sqrt{S^2}$ ist nicht erwartungstreu für $\sqrt{\sigma^2} = \sigma$.

Bsp. 2.6. [Wahlumfrage]

Gegeben sei eine Stichprobe der wahlberechtigten Bundesbürger. Geben Sie einen erwartungstreuen Schätzer des Anteils der rot-grün Wähler an.

Bedeutung der Erwartungstreue: Erwartungstreue ist in gewisser Weise ein schwaches Kriterium, denn es gibt viele unsinnige erwartungstreue Schätzer!

Deshalb betrachtet man zusätzlich die Effizienz eines Schätzers, s.u.

2.2.3 Effizienz

Beispiel Wahlumfrage: Gegeben sind zwei erwartungstreue Schätzer (n sei zum einfacheren Rechnen gerade):

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$T_2 = \frac{1}{n/2}\sum_{i=1}^{n/2} X_i$$

Was unterscheidet formal T_1 von dem unsinnigen Schätzer T_2, der die in der Stichprobe enthaltene Information nicht vollständig ausnutzt?

Vergleiche die Schätzer über ihre Varianz, nicht nur über den Erwartungswert!

Wenn n so groß ist, dass der zentrale Grenzwertsatz angewendet werden kann, dann gilt approximativ

$$\frac{1}{\sqrt{n}}\frac{\sum_{i=1}^{n}(X_i-\pi)}{\sqrt{\pi(1-\pi)}}=\frac{\sum_{i=1}^{n}X_i-n\cdot\pi}{\sqrt{n}\sqrt{\pi(1-\pi)}}=\frac{\frac{1}{n}\sum_{i=1}^{n}X_i-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}\sim\mathcal{N}(0;1)$$

und damit

$$T_1=\frac{1}{n}\sum_{i=1}^{n}X_i\sim\mathcal{N}\left(\pi;\frac{\pi(1-\pi)}{n}\right).$$

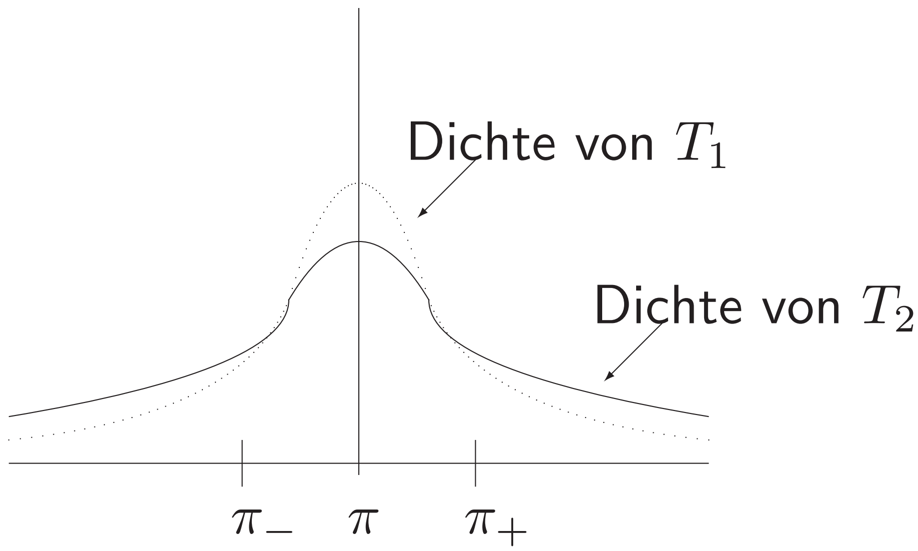
Analog kann man zeigen:

$$T_2=\frac{1}{n/2}\sum_{i=1}^{n/2}X_i\sim\mathcal{N}\left(\pi,\frac{\pi(1-\pi)}{n/2}\right).$$

T_1 und T_2 sind approximativ normalverteilt, wobei T_1 eine deutlich kleinere Varianz als T_2 hat.

T_1 und T_2 treffen beide im Durchschnitt den richtigen Wert π. Der Schätzer T_1 schwankt aber weniger um das wahre π, ist also „im Durchschnitt genauer“.

Andere Interpretation:



Für jeden Punkt $\pi_+ > \pi$ ist damit $P(T_1 > \pi_+) < P(T_2 > \pi_+)$
und für jeden Punkt $\pi_- < \pi$ ist $P(T_1 < \pi_-) < P(T_2 < \pi_-)$.

Es ist also die Wahrscheinlichkeit, mindestens um $\pi_+ - \pi$ bzw. $\pi - \pi_-$ daneben zu liegen, bei T_2 stets größer als bei T_1. Umgekehrt gesagt: Ein konkreter Wert ist damit verlässlicher, wenn er von T_1, als wenn er von T_2 stammt.

Diese Überlegung gilt ganz allgemein: Ein erwartungstreuer Schätzer ist umso besser, je kleiner seine Varianz ist.

$$\mathrm{Var}(T) = \text{Erwartete quadratische Abweichung von } T \text{ von } \underbrace{\mathrm{E}(T)}_{=\vartheta}$$

Je kleiner die Varianz, umso mehr konzentriert sich die Verteilung eines erwartungstreuen Schätzers um den wahren Wert. Dies ist umso wichtiger, da der Schätzer den wahren Wert i.A. nur selten exakt trifft.

Definition 2.7. *Effizienz*

- Gegeben seien zwei erwartungstreue Schätzfunktionen T_1 und T_2 für einen Parameter ϑ. Gilt

$$\mathrm{Var}_\vartheta(T_1) \leq \mathrm{Var}_\vartheta(T_2) \text{ für alle } \vartheta$$

und

$$\mathrm{Var}_{\vartheta^*}(T_1) < \mathrm{Var}_{\vartheta^*}(T_2) \text{ für mindestens ein } \vartheta^*$$

so heißt T_1 *effizienter als* T_2.

- Eine für ϑ erwartungstreue Schätzfunktion T^{opt} heißt *UMVU-Schätzfunktion* für ϑ ($u$niformly $m$inimum $v$ariance $u$nbiased), falls

$$\mathrm{Var}_\vartheta(T^{\text{opt}}) \leq \mathrm{Var}_\vartheta(T)$$

für alle ϑ und für alle erwartungstreuen Schätzfunktionen T.

Bem. 2.8.

- *Inhaltliche Bemerkung:* Der „Sinn von Optimalitätskriterien“ bei der Auswertung der Stichprobe wird klassischerweise insbesondere auch in der „*Gewährleistung von Objektivität*“, besser Intersubjektivität, gesehen. Ohne wissenschaftlichen Konsens darüber, welcher Schätzer in welcher Situation zu wählen ist, wäre die Auswertung einer Stichprobe willkürlich und der Manipulation Tür und Tor geöffnet. Allerdings gibt es wirkliche Eindeutigkeit nur bei „idealen“, sauberen Daten. Z.B. sind ausreißerunempfindliche Verfahren bei „idealen“ Daten weniger effizient, haben aber den Vorteil, stabiler bei kleinen Abweichungen von den Verteilungsannahmen zu sein.

- Ist $X_1, \ldots, X_n$ eine i.i.d. Stichprobe mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$, dann ist

  * $\bar{X}$ UMVU-Schätzfunktion für μ bei bekanntem σ^2 und

  * S^2 UMVU-Schätzfunktion für σ^2 bei bekanntem μ.

- Ist $X_1, \ldots, X_n$ mit $X_i \in \{0, 1\}$ eine i.i.d. Stichprobe mit $\pi = P(X_i = 1)$, dann ist die relative Häufigkeit $\bar{X}$ UMVU-Schätzfunktion für π.

- Bei nicht erwartungstreuen Schätzern macht es keinen Sinn, sich ausschließlich auf die Varianz zu konzentrieren.

  Man zieht dann den sogenannten *Mean Squared Error*

$$\mathrm{MSE}_\vartheta(T) \quad := \quad \mathrm{E}_\vartheta(T-\vartheta)^2$$

  zur Beurteilung heran. Es gilt

$$\mathrm{MSE}_\vartheta(T) \quad = \quad \mathrm{Var}_\vartheta(T) + (\mathrm{Bias}_\vartheta(T))^2.$$

  Der MSE kann als Kompromiss zwischen zwei Auffassungen von Präzision gesehen werden: möglichst geringe systematische Verzerrung (Bias) und möglichst geringe Schwankung (Varianz).

2.2.4 Asymptotische Gütekriterien

- **Asymptotische Erwartungstreue**

  * Eine Schätzfunktion heißt *asymptotisch erwartungstreu*, falls

  $$\lim_{n \to \infty} E_\vartheta(\hat{\theta}) = \theta$$

  d.h.,

  $$\lim_{n \to \infty} \mathrm{Bias}_\vartheta(\hat{\theta}) = 0$$

  gilt.

  * Abschwächung des Begriffs der Erwartungstreue: Gefordert wird die Erwartungstreue nur bei einer unendlich großen Stichprobe.

  * Erwartungstreue Schätzer sind auch asymptotisch erwartungstreu.

  * Sowohl S^2 als auch $\tilde{S}^2$ sind asymptotisch erwartungstreu.
  (vgl. vorne $\mathrm{Bias}_{\sigma^2}(\tilde{S}^2) = -\frac{1}{n}\sigma^2$ geht für $n \to \infty$ gegen 0.)

- Für komplexere Modelle ist oft die Erwartungstreue der Verfahren ein zu restriktives Kriterium. Man fordert deshalb oft nur, dass sich der Schätzer wenigstens für große Stichproben gut verhält. Hierzu gibt es verwandte aber „etwas“ unterschiedliche Kriterien, z.B. das folgende:

- Ein Schätzer heißt *(MSE-)konsistent* oder *konsistent im quadratischen Mittel*, wenn gilt

$$\lim_{n\to\infty}(\mathrm{MSE}(T)) = 0.$$

Beispiel: Der MSE von $\bar{X}$ ist gegeben durch

$$\begin{aligned}\mathrm{MSE}(\bar{X}) &= \mathrm{Var}(\bar{X}) + \mathrm{Bias}^2(\bar{X})\\ &= \frac{\sigma^2}{n} + 0\\ &= \frac{\sigma^2}{n} \to 0.\end{aligned}$$

$\bar{X}$ ist also ein MSE-konsistenter Schätzer für den Erwartungswert.

- Anschaulich bedeutet die Konsistenz,

2.2.5 Konstruktionsprinzipien guter Schätzer

- **Die Methode der kleinsten Quadrate**
  $\Longrightarrow$ Regressionsanalyse

- **Das Maximum-Likelihood-Prinzip**

* Aufgabe: Schätze den Parameter ϑ eines parametrischen Modells anhand einer i.i.d. Stichprobe $X_1, \dots, X_n$ mit der konkreten Realisation $x_1, \dots, x_n$.

* Idee der Maximium-Likelihood (ML) Schätzung für diskrete Verteilungen:

- Man kann für jedes ϑ die Wahrscheinlichkeit ausrechnen, genau die Stichprobe

$x_1, \dots, x_n$ zu erhalten:

$$P_\vartheta(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^{n} P_\vartheta(X_i = x_i)$$

- Je größer für ein gegebenes ϑ_0 die Wahrscheinlichkeit ist, die konkrete Stichprobe erhalten zu haben, umso plausibler ist es, dass tatsächlich ϑ_0 der wahre Wert ist (gute Übereinstimmung zwischen Modell und Daten).
- Man nennt daher $L(\vartheta) = P_\vartheta(X_1 = x_1, \dots, X_n = x_n)$, nun als Funktion von ϑ gesehen, die *Likelihood* (deutsch: Plausibilität, Mutmaßlichkeit) von ϑ gegeben die Realisation $x_1, \dots, x_n$.
- Derjenige Wert $\hat{\vartheta} = \hat{\vartheta}(x_1, \dots, x_n)$, der $L(\vartheta)$ maximiert, heißt *Maximum-Likelihood-Schätzwert*; die zugehörige Schätzfunktion $T(X_1, \dots, X_n)$ *Maximum-Likelihood-Schätzer* (siehe genauer Definition 2.11).

Bsp. 2.9.

i.i.d. Stichprobe vom Umfang $n = 5$ aus einer $B(10, \pi)$-Verteilung:

6 5 3 4 4

Wahrscheinlichkeit der Stichprobe für gegebenes π, hier notiert durch „$||\ \pi$“:
„$P(\ldots||\pi)$ Wahrscheinlichkeit, wenn π der wahre Parameter ist“

$$\begin{aligned} P(X_1 = 6, \ldots, X_5 = 4 \,||\, \pi) &= P(X_1 = 6 \,||\, \pi) \cdot \ldots \cdot P(X_5 = 4 \,||\, \pi) \\ &= \binom{10}{6} \pi^6 (1-\pi)^4 \cdot \ldots \cdot \binom{10}{4} \pi^4 (1-\pi)^6. \end{aligned}$$

Wahrscheinlichkeit für einige Werte von π:

π	$P(X_1 = 6, \ldots, X_5 = 4 \,\|\|\, \pi)$
0.1	0.0000000000001
0.2	0.0000000227200
0.3	0.0000040425220
0.4	0.0003025481000
0.5	0.0002487367000
0.6	0.0000026561150
0.7	0.0000000250490
0.8	0.0000000000055
0.9	0.0000000000000

Bem. 2.10.

- Zwei Sichtweisen auf
$$P_\vartheta(X_1 = x_1, \ldots, X_n = x_n):$$
  * Deduktiv (Wahrscheinlichkeitsrechnung): ϑ bekannt, $x_1, \ldots, x_n$ zufällig („unbekannt“).
  * Induktiv (Statistik): ϑ unbekannt, $x_1, \ldots, x_n$ bekannt.
- Für stetige Verteilungen gilt
$$P_\vartheta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = 0$$
für beliebige Werte ϑ. In diesem Fall verwendet man die Dichte
$$f_\vartheta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\vartheta(x_i)$$
als Maß für die Plausibilität von ϑ.

- Für die praktische Berechnung maximiert man statt der Likelihood typischerweise die sog. *Log-Likelihood* $l(\vartheta)$, also den natürlichen Logarithmus der Likelihood.

$$l(\vartheta) = \ln(L(\vartheta)) = \ln \prod_{i=1}^{n} P_\vartheta(X_i = x_i) = \sum_{i=1}^{n} \ln P_\vartheta(X_i = x_i)$$

bzw.

$$l(\vartheta) = \ln \prod_{i=1}^{n} f_\vartheta(x_i) = \sum_{i=1}^{n} \ln f_\vartheta(x_i).$$

Dies liefert denselben Schätzwert $\hat{\vartheta}$ und erspart beim Differenzieren die Anwendung der Produktregel. Manchmal ist es noch geschickter, zunächst $\prod_{i=1}^{n} P_\vartheta(X_i = x_i)$ zu vereinfachen und dann zu logarithmieren.

- Bei den in Statistik II betrachteten „regulären“ Verteilungsmodellen reicht es, die erste Ableitung zu betrachten. Man kann zeigen, dass sie immer auf ein Maximum führt.

Definition 2.11. *(Maximum-Likelihood-Schätzung)*

Gegeben sei die Realisation $x_1, \ldots, x_n$ einer i.i.d. Stichprobe. Die Funktion in ϑ

$$L(\vartheta) = \begin{cases} \prod_{i=1}^{n} P_\vartheta(X_i = x_i) & \text{falls } X_i \text{ diskret} \\ \prod_{i=1}^{n} f_\vartheta(x_i) & \text{falls } X_i \text{ stetig.} \end{cases}$$

heißt *Likelihood* des Parameters ϑ bei der Beobachtung $x_1, \ldots, x_n$.

Derjenige Wert $\hat{\vartheta} = \hat{\vartheta}(x_1, \ldots, x_n)$, der $L(\vartheta)$ maximiert, heißt *Maximum-Likelihood-Schätzwert*; die zugehörige Schätzfunktion $T(X_1, \ldots, X_n)$ *Maximum-Likelihood-Schätzer*.

Bsp. 2.12. *(Maximum-Likelihood-Schätzer bei der Bernoulli-Verteilung)*

Bsp. 2.13. [ML-Schätzung bei Normalverteilung]

- Der ML-Schätzer $\hat{\mu} = \bar{X}$ für μ stimmt mit dem üblichen Schätzer für den Erwartungswert überein.

- Der ML-Schätzer $\hat{\sigma}^2 = \tilde{S}^2$ für σ^2 ist die Stichprobenvarianz; diese ist nicht erwartungstreu.

Bem. 2.14. [Einige allgemeine Eigenschaften von ML-Schätzern]

Unter allgemeinen Regularitätsbedingungen gilt:

- ML-Schätzer $\hat{\theta}$ sind im Allgemeinen nicht erwartungstreu.
- ML-Schätzer $\hat{\theta}$ sind asymptotisch erwartungstreu.
- ML-Schätzer $\hat{\theta}$ sind konsistent (und meist in einem asymptotischen Sinne effizient).

2.3 Intervallschätzung

2.3.1 Motivation und Hinführung

Bsp. 2.15. [Wahlumfrage]

Der wahre Anteil der „rot-grün Wähler“ unter allen Wählern war 2009 –auf eine Nachkommastelle gerundet– genau 33.7%. Wie groß ist die Wahrscheinlichkeit, in einer Zufallsstichprobe von 1000 Personen aus allen Wählern genau einen relativen Anteil von 33.7% von „rot-grün Wählern“ erhalten zu haben?

D.h., mit Wahrscheinlichkeit von etwa 97.3%, verfehlt der Schätzer den wahren Wert.

Hat man ein stetiges Merkmal, so ist sogar $P(X = w) = 0$ für jeden Wert w, d.h. der wahre Wert wird mit Wahrscheinlichkeit 1 verfehlt.

Konsequenzen:

- Insbesondere Vorsicht bei der Interpretation „knapper Ergebnisse“ (z.B. Anteil 50.2%)
- Suche Schätzer mit möglichst kleiner Varianz, um „im Durchschnitt möglichst nahe am wahren Wert dran zu sein“
- Es ist häufig auch gar nicht nötig, sich genau auf einen Wert als Schätzung festzulegen. Oft reicht die Angabe eines Intervalls, von dem man hofft, dass es den wahren Wert überdeckt: *Intervallschätzung*

Symmetrische Intervallschätzung basierend auf einer Schätzfunktion $T = g(X_1, \ldots, X_n)$:

$$I(T) = [T - a, T + a]$$

„Trade off“ bei der Wahl von a:

$\Rightarrow$ Wie soll man a wählen?

Typisches Vorgehen:

- Man gebe sich durch inhaltliche Überlegungen einen Sicherheitsgrad (*Konfidenzniveau*) γ vor.
- Dann konstruiert man das Intervall so, dass es mindestens mit der Wahrscheinlichkeit γ den wahren Parameter überdeckt $\Rightarrow$ „Konfidenzintervall“ („Vertrauensintervall“)

2.3.2 Definition von Konfidenzintervallen:

Gegeben sei eine i.i.d. Stichprobe $X_1, \ldots, X_n$ zur Schätzung eines Parameters ϑ und eine Zahl $\gamma \in (0; 1)$. Ein zufälliges Intervall $\mathcal{C}(X_1, \ldots, X_n)$ heißt *Konfidenzintervall* zum *Sicherheitsgrad* γ (Konfidenzniveau γ), falls für jedes ϑ gilt:

$$P_\vartheta(\vartheta \in \underbrace{\mathcal{C}(X_1, \ldots, X_n)}_{\text{zufälliges Intervall}}) \geq \gamma.$$

Bem. 2.16.

- Die Wahrscheinlichkeitsaussage bezieht sich auf das Ereignis, dass das zufällige Intervall den festen, wahren Parameter überdeckt. Streng genommen darf man im objektivistischen Verständnis von Wahrscheinlichkeit nicht von der $Wahrscheinlichkeit$ sprechen, „dass ϑ in dem Intervall liegt“, da ϑ nicht zufällig ist und somit keine Wahrscheinlichkeitsverteilung besitzt. (Ein Zurückziehen auf subjektive Wahrscheinlichkeiten führt hier aber auch auf Widersprüche.)

- Typischerweise konstruiert man Konfidenzintervalle symmetrisch um einen Schätzer T. Es sind aber auch manchmal z.B. einseitige Konfidenzintervalle, etwa der Form

$$[\bar{X}, \bar{X} + b],$$

sinnvoll.

2.3.3 Konstruktion von Konfidenzintervallen

Für die Konstruktion sehr praktische Vorgehensweise: Suche Zufallsvariable Z_ϑ, die

- den gesuchten Parameter ϑ enthält und
- deren Verteilung aber nicht mehr von dem Parameter abhängt, („*Pivotgröße*“, dt. Angelpunkt).
- Dann wähle den Bereich $[z_{\text{unten}}, z_{\text{oben}}]$ so, dass $P(Z_\vartheta \in [z_{\text{unten}}, z_{\text{oben}}]) = \gamma$ und
- löse nach ϑ auf.

Mit Blick auf Kapitel 1.5.4 kann man also sagen: Man bildet für Z_ϑ einen Modalbereich zum Grade γ und löst dann nach ϑ auf.

Konfidenzintervall für den Mittelwert eines normalverteilten Merkmals bei bekannter Varianz:

$X_1, \ldots, X_n$ i.i.d. Stichprobe gemäß $X_i \sim \mathcal{N}(\mu, \sigma^2)$, wobei σ^2 bekannt sei.

$$\left[\bar{X} - \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}}\right] = \left[\bar{X} \pm \frac{z_{\frac{1+\gamma}{2}} \cdot \sigma}{\sqrt{n}}\right] \tag{2.10}$$

Bem. 2.17.

i) Die Wirkung der einzelnen Größen im Konfidenzintervall erkennt man am besten durch eine „*Ceteris paribus-Analyse*“; alle Größen bis auf eine werden festgehalten, diese wird variiert:

   – Je größer σ, desto breiter das Intervall!

   – Je größer γ, desto größer $z_{\frac{1+\gamma}{2}}$

   – Je größer n und damit $\sqrt{n}$, desto schmäler ist das Intervall

ii) Man vergleicht dies mit dem entsprechenden Modalbereich am Ende von Kap. 1.6 mit X durch $\bar{X}$ und damit σ^2 durch $\frac{\sigma^2}{n}$ ersetzt

Konfidenzintervall für den Mittelwert eines normalverteilten Merkmals bei unbekannter Varianz:

Neben dem Erwartungswert sei nun auch σ^2 unbekannt. σ^2 wird entsprechend durch den UMVU-Schätzer

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

(mit $S = \sqrt{S^2}$) geschätzt. Allerdings ist

$$Z = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

jetzt nicht mehr normalverteilt, denn S ist zufällig.

Wir führen deshalb ein neues Verteilungsmodell ein.

Bem. 2.18. *t-Verteilung*

Gegeben sei eine i.i.d. Stichprobe $X_1, \ldots, X_n$ mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Dann heißt die Verteilung von

$$Z = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

t-$Verteilung$ (oder Student-Verteilung) mit $\nu = n - 1$ $Freiheitsgraden$. In Zeichen: $Z \sim t(\nu)$.

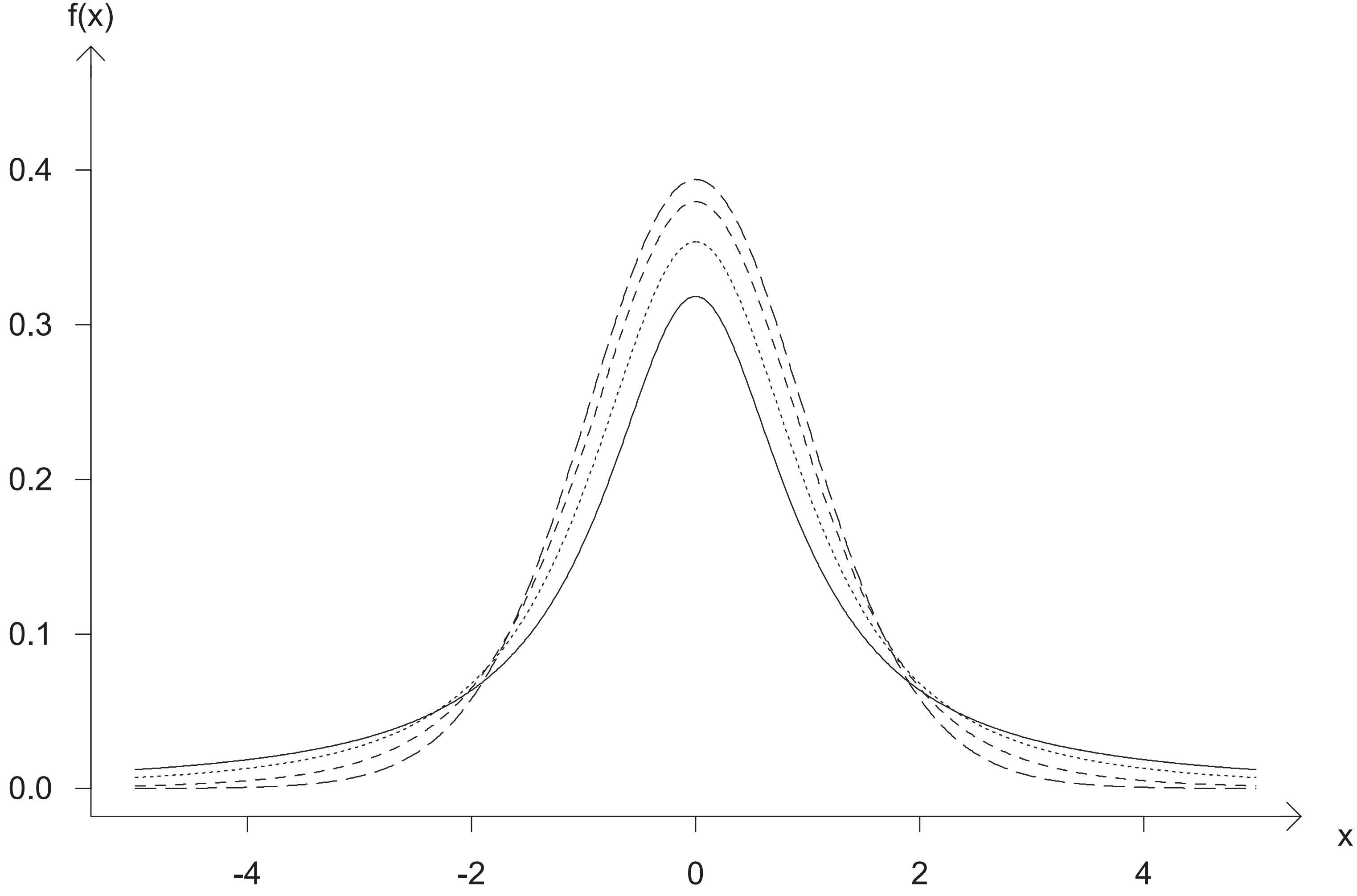
Wichtige Werte der t-Verteilung sind tabelliert. Angegeben sind, für verschiedene α, die Lösung t_α der Gleichung

$$P(Z \leq t_\alpha(\nu)) = \alpha,$$

wobei $t_\alpha(\nu)$ von der Anzahl ν der Freiheitsgrade abhängt. t_α ist das α-Quantil der entsprechenden t-Verteilung (analog zu z_α als Quantil der Standardnormalverteilung).

Die Dichte einer t-Verteilung ist der Dichte der Standardnormalverteilung sehr ähnlich: Sie ist auch symmetrisch um 0, besitzt aber etwas höhere Dichte für extreme Werte („schwerere Enden“).

ν	0.8	0.9	0.95	0.975	0.99	0.995	0.999	0.9995
1	1.3764	3.0777	6.3138	12.706	31.821	63.657	318.31	636.62
2	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248	22.327	31.599
3	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409	10.215	12.924
4	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103
5	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688
6	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588
7	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079
8	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413
9	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809
10	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869
11	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370
12	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178
13	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208
14	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405
15	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728
16	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862	4.0150
17	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458	3.9651
18	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105	3.9216
19	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794	3.8834
20	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518	3.8495
∞	0.8416	1.2816	1.6449	1.9600	2.3263	2.5758	3.0903	3.2906



Dichten von t-Verteilungen für $\nu = 1$ (—), $= 2$ $(\cdots)$, $= 5$ (- - -) und $= 20$ $(——)$ Freiheitsgrade.

Je größer ν ist, umso ähnlicher sind sich die $t(\nu)$-Verteilung und die Standardnormalverteilung. Für $\nu \to \infty$ sind sie gleich, ab $\nu = 30$ gilt im Allgemeinen der Unterschied als vernachlässigbar.

Konfidenzintervall zum Konfidenzniveau γ:

Bem. 2.19.

- Es gelten analoge Aussagen zur Wirkung des Stichprobenumfangs und Konfidenzniveaus wie bei bekannter Varianz. (vgl. Bem. 2.17)
- Für jedes γ (und jedes ν) gilt
$$t_{\frac{1+\gamma}{2}}(\nu) > z_{\frac{1+\gamma}{2}} \tag{2.11}$$
also ist das t-Verteilungs-Konfidenzintervall (etwas) breiter.
- Da die Unterschiede zur Normalverteilung für $n \geq 30$ meist als vernachlässigbar gelten (s.o.), rechnet man für $n \geq 30$ bei der t-Verteilung auch einfach mit $z_{\frac{1+\gamma}{2}}$.

Bsp. 2.20.

Eine Maschine füllt Gummibärchen in Tüten ab, die laut Aufdruck 250g Füllgewicht versprechen. Wir nehmen im Folgenden an, dass das Füllgewicht normalverteilt ist. Bei 16 zufällig aus der Produktion herausgegriffenen Tüten wird ein mittleres Füllgewicht von 245g und eine Stichprobenstreuung (Standardabweichung) von 10g festgestellt.

a) Berechnen Sie ein Konfidenzintervall für das mittlere Füllgewicht zum Sicherheitsniveau von 95%.

b) Wenn Ihnen zusätzlich bekannt würde, dass die Stichprobenstreuung gleich der tatsächlichen Streuung ist, wäre dann das unter a) zu berechnende Konfidenzintervall für das mittlere Füllgewicht breiter oder schmäler? Begründen Sie ihre Antwort ohne Rechnung!

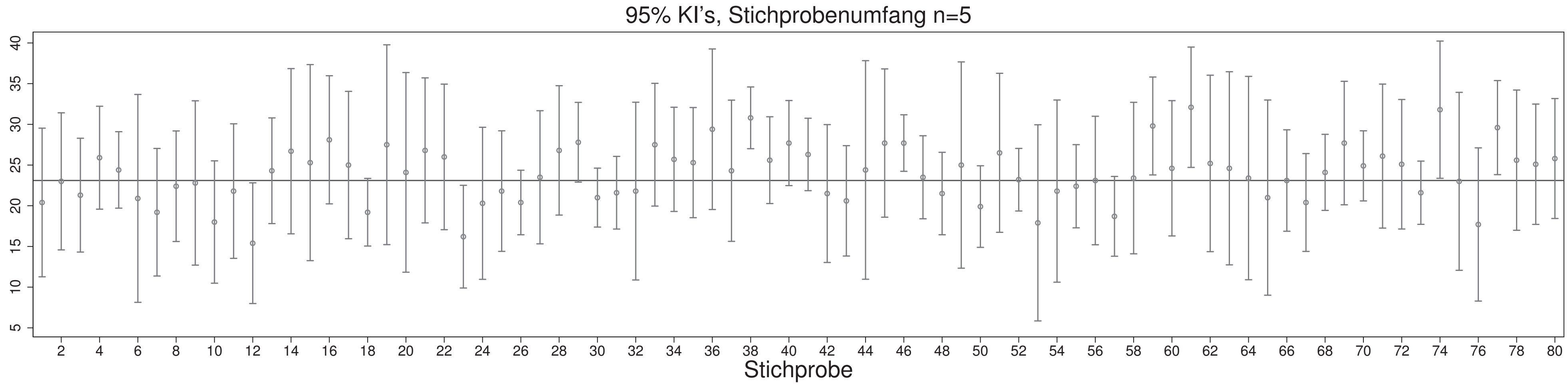
Bsp. 2.21. [Klausurergebnisse (Fiktives Beispiel)]

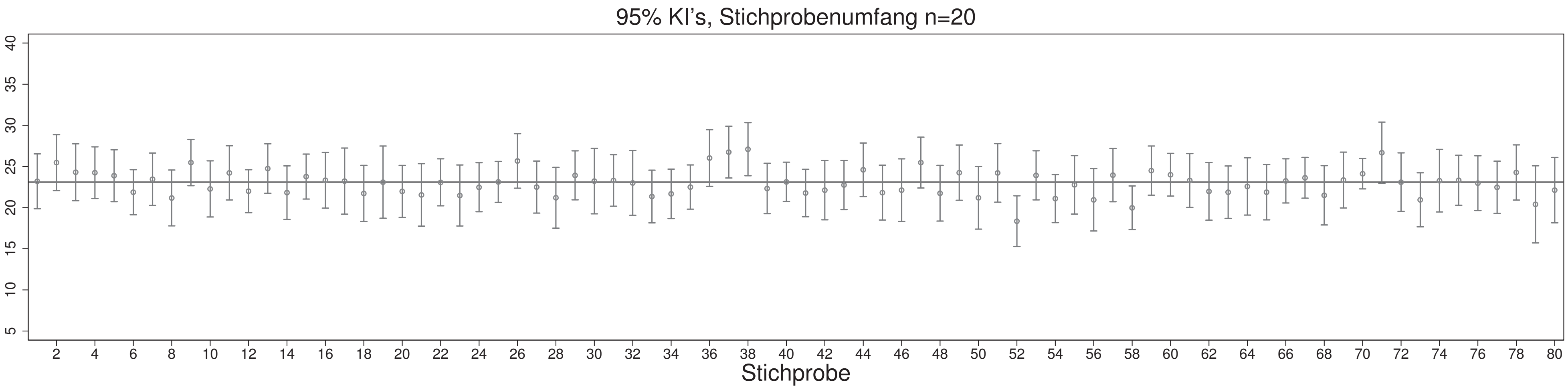
i	punkte	i	punkte	i	punkte	i	punkte	i	punkte	i	punkte
1	16.5	7	27.5	13	23	19	14	25	19.5	31	35.5
2	16.5	8	22	14	15	20	21	26	18	32	12.5
3	25.5	9	16.5	15	31	21	19.5	27	37.5	33	25
4	25	10	8	16	26	22	17.5	28	15.5	34	25.5
5	20.5	11	33.5	17	13.5	23	36	29	7.5	35	18.5
6	27.5	12	19.5	18	24	24	31.5	30	18		

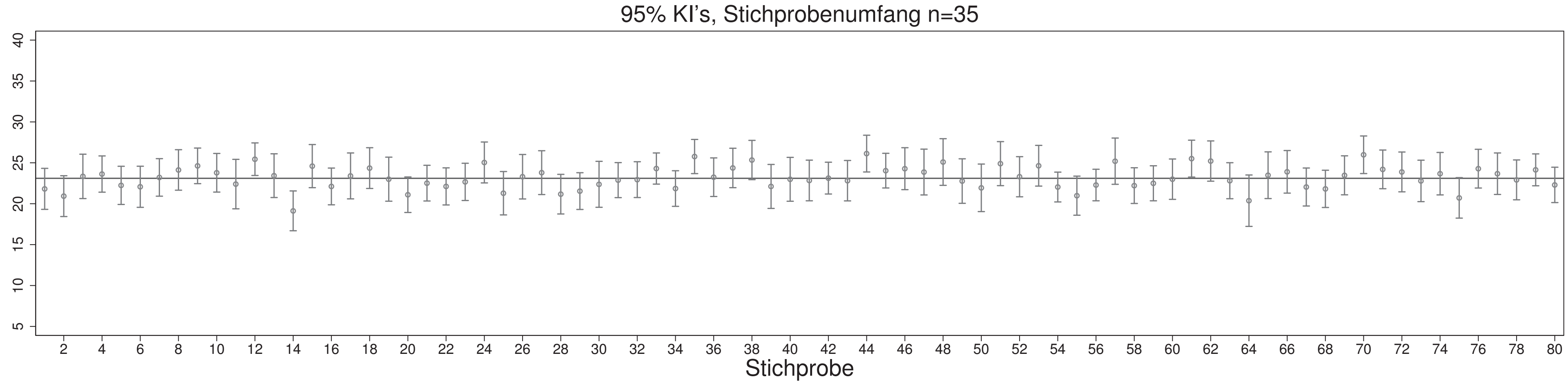
- Mittelwert und Varianz der Grundgesamtheit (alle 35 Klausuren):

$$\begin{aligned} \mu &= \overline{\text{punkte}} = 21.81 \\ \sigma^2 &= s^2_{\text{punkte}} = 7.56^2 \end{aligned}$$

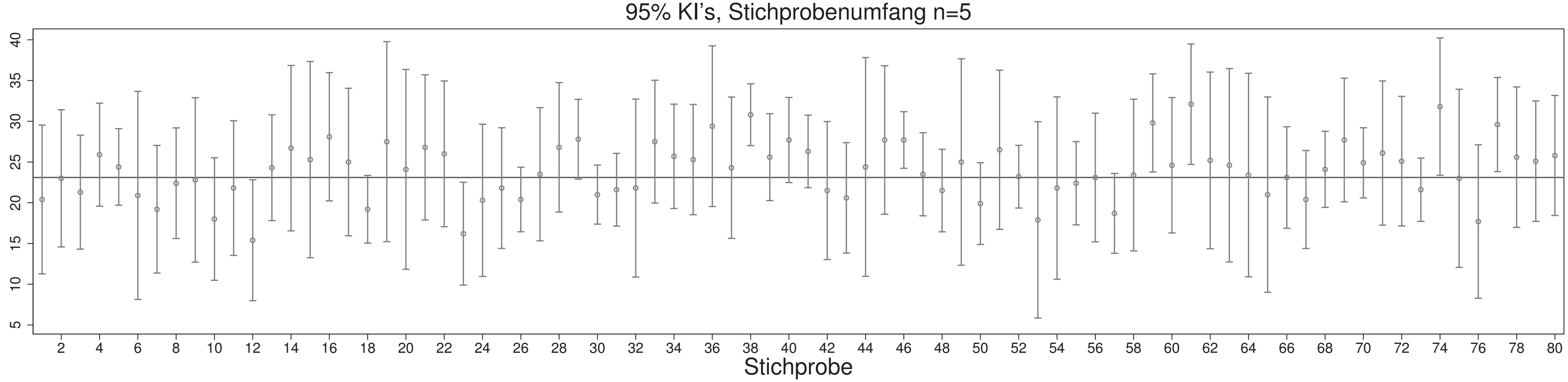
- Wir ziehen insgesamt jeweils 80 Stichproben vom Umfang $n = 5$, $n = 20$, $n = 35$ und bestimmen Konfidenzintervalle zum Niveau 95%.

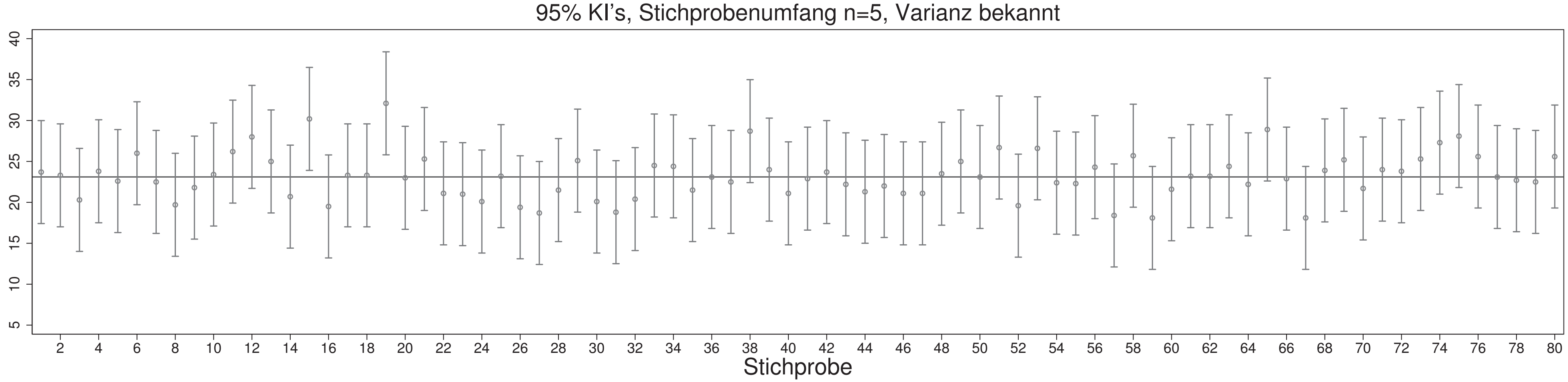

95% KI's, Stichprobenumfang n=5
5
10
15
20
25
30
35
40
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80
Stichprobe


95% KI's, Stichprobenumfang n=20
5
10
15
20
25
30
35
40
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80
Stichprobe



Punktzahl der Klausur Statistik I: Konfidenzintervalle für 80 Stichproben mit jeweiligem Stichprobenumfang $n = 5$ (oben), $n = 20$ (Mitte), $n = 35$ (unten).





Punktzahl der Klausur Statistik I: Konfidenzintervalle für 80 Stichproben mit Stichprobenumfang $n = 5$. Oben ist die Varianz unbekannt, unten als bekannt vorausgesetzt.

- Ergebnis:

  * Die Breite der Konfidenzintervalle nimmt mit wachsendem n ab.
  * Nicht alle Konfidenzintervalle enthalten den wahren Mittelwert $\mu = 23.1$ (per Konstruktion tritt dieser Effekt mit Wahrscheinlichkeit 5% ein).
  * Die Intervalle mit bekannter Varianz sind im Mittel enger.
  * Die Intervalle mit bekannter Varianz sind immer gleich lang.

2.3.4 Approximative Konfidenzintervalle

Ist der Stichprobenumfang groß genug, so kann wegen des zentralen Grenzwertsatzes das Normalverteilungs-Konfidenzintervall auf den Erwartungswert beliebiger Merkmale (mit existierender Varianz) angewendet werden. Man erhält approximative Konfidenzintervalle, die meist auch der Berechnung mit Software zugrundeliegen.

Beispiel: Konfidenzintervall für den Anteil π: Seien $X_1, \ldots, X_n$ i.i.d. mit

$$X_i = \begin{cases} 1 \\ 0 \end{cases}, \quad P(X_i = 1) = \pi.$$

Bsp. 2.22. [Wahlumfrage]

Seien $n = 500$, $\bar{X} = 46.5\%$ und $\gamma = 95\% \Rightarrow \frac{1+\gamma}{2} = 97.5\%$.

Inhaltliche Bemerkungen:

2.3.5 Bestimmung des Stichprobenumfangs

Eine große praktische Bedeutung haben Konfidenzintervalle auch für die Stichprobenplanung. Sie werden herangezogen, um den benötigten Stichprobenumfang zu ermitteln.

- Je genauer und sicherer, desto größer muss der Stichprobenumfang sein
- Genauigkeit: Halbe Länge des Konfidenzintervalls
- Gib Konfidenzniveau (oft 95%) und eine Schranke b_{max} an (inhaltliche Entscheidungen!), also de facto die halbe Breite für die minimal zulässige Genauigkeit im Sinne von $\frac{1}{b_{\text{max}}}$
- Bestimme den Mindeststichprobenumfang so, dass γ und b_{max} eingehalten werden

Konkrete Umsetzung für die Anteilsschätzung

γ: Konfidenzniveau
b_{max}: Genauigkeitsschranke

$$z_{\frac{1+\gamma}{2}} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \quad \leq \quad b_{\text{max}}$$

Auflösen nach n:

$$n \quad \geq \quad \frac{1}{b_{\text{max}}^2} z_{\frac{1+\gamma}{2}}^2 \cdot \bar{X}(1-\bar{X})$$

Da $\bar{X}$ unbekannt ist, verwendet man zur Stichprobenplanung z.B. Schätzwerte früherer Untersuchungen oder als konservative Schranke $\bar{X}(1-\bar{X}) \leq 0.25$.[3]

[3]Dies gilt immer:
betrachte $f(a) = a \cdot (1-a) = -a^2 + a;\; a \in [0,1]$
$\frac{df(a)}{d(a)} = 0 \Leftrightarrow -2a + 1 = 0 \Rightarrow a = \frac{1}{2}$
$\frac{d^2 f(a)}{da^2} = -2$
also Maximum an der Stelle $a_{\text{max}} = \frac{1}{2}$ mit $f(a)_{\text{max}} = (-\frac{1}{2})^2 + \frac{1}{2} = \frac{1}{4}$

Bsp. 2.23.

Gegeben:

- Konfidenzniveau: $\gamma = 95\%$
- Genauigkeit: $b_{\text{max}} = 0.1$

Bestimmung von n:

$$n \geq \frac{1}{g^2} z^2_{\frac{1+\gamma}{2}} \cdot \bar{X}(1 - \bar{X}) = \frac{1}{0.1^2} 1.96^2 \cdot 0.25 = 96.04$$

wobei konservativ angesezt wurde $\bar{X}(1 - \bar{X}) \leq 0.25$. Also sollten ca. 100 Personen befragt werden.

Bei $g = 5\%$ ergibt sich $n = 385$, bei $g = 1\%$ ergibt sich $n = 9604$

2.3.6 Ausblick

Viele weitere Konfidenzintervalle, auch bezüglich der Unterschiede bei zwei Stichproben

- Differenz von Mittelwerten bei unabhängigen Stichproben
- Differenz von Anteilen bei unabhängigen Stichproben
- Differenz von Mittelwerten bei „verbundenen Stichproben“

Diese gehorchen immer dem gleichen Konstruktionsprinzip (vgl. von Kap. 2.3.3) und sind analog zu den genauer besprochenenen Beispielen zu interpretieren. Man vergleiche auch die Konstruktion von Tests im Kapitel **??**, die auf einem ähnlichen Ausgangspunkt rekurrieren. In der Tat gibt es einen engen Zusammenhang zwischen Hypothesentests und Konfidenzintervallen (vgl. „Dualität“ am Ende von Kap. **??**).